DOCUMENT RESUME

ED 340 753 TM 018 009

AUTHOR Banta, Trudy W.
TITLE Toward a Plan for Using National Assessment To Ensure Continuous Improvement of Higher Education. Draft.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
PUB DATE 30 Sep 91
NOTE 50p.; Commissioned paper prepared for a workshop on Assessing Higher Order Thinking & Communication Skills in College Graduates (Washington, DC, November 17-19, 1991), in support of National Education Goal V, Objective 5. For other workshop papers, see TM 018 010-024.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *College Graduates; *Communication Skills; *Critical Thinking; *Educational Assessment; Educational Improvement; Educational Objectives; Evaluation Utilization; Higher Education; Measurement Techniques; National Programs; Outcomes of Education; *Problem Solving; Student Evaluation; Testing Programs; *Thinking Skills
IDENTIFIERS America 2000; *National Education Goals 1990

ABSTRACT

An outline is provided for a national educational assessment and improvement plan as suggested by the National Education Goals of 1990. The following implicit assumptions underlie National Education Objective 5.5: (1) abilities to think critically, communicate effectively, and solve problems can be defined and the definitions can be agreed upon as desired instructional objectives; (2) defined abilities can be taught in ways that engage students and promote learning; (3) reliable and valid measures of these abilities can be identified and created; (4) measures of student attainment can be administered to college graduates in settings that encourage their best efforts; and (5) results of such assessment will be used to improve instruction. Under prevailing conditions in American higher education, little support for these assumptions exists. Current measurement theory is inadequate to provide direction for teaching and learning, and the act of assessing student abilities will not, in and of itself, improve those abilities. If decision makers believe that the national interest will be best served by a comprehensive postsecondary assessment program, the principles of continuous improvement applied in industry should be used to link faculty goal-setting, staff development, assessment of instructional resources and student outcomes, and uses of assessment results for educational improvement. A 55-item list of references is included. Reviews by N. Frederiksen and by B. Wright and T. Marchese of this position paper are provided. (SLD)

ED340753

DRAFT

September 30, 1991

# TOWARD A PLAN FOR USING NATIONAL ASSESSMENT TO ENSURE CONTINUOUS IMPROVEMENT OF HIGHER EDUCATION

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

Paper prepared for

The National Center for Education Statistics

By:

Trudy W. Banta
Professor and Director

Center for Assessment Research and Development
1819 Andy Holt Avenue
Knoxville, Tennessee 37996-4350
(615) 974-2350

TM018009



BEST COPY AVAILABLE

# TOWARD A PLAN FOR USING NATIONAL ASSESSMENT TO ENSURE CONTINUOUS IMPROVEMENT OF HIGHER EDUCATION

Trudy W. Banta

– Abstract –

Underlying National Education Objective 5.5 are several implicit assumptions. Five of these, which are considered in some detail in this paper, may be stated as follows: 1) The abilities to "think critically, communicate effectively, and solve problems" can be operationally defined and these definitions agreed upon as desired instructional outcomes by all U.S. faculty responsible for developing these abilities in undergraduates; 2) The defined abilities will be taught, by all faculty charged with the responsibility for teaching them, in ways that engage students and promote learning of these abilities; 3) Reliable and valid measures of student achievement of the defined abilities can be identified or created; 4) The measures of student attainment can be administered to college graduates in settings that engage students and encourage their best efforts; 5) The results of assessment of developed student abilities will be used to improve the materials and methods of instruction in ways that increase student engagement and promote learning gains.

Under prevailing conditions in American higher education, little evidence exists to support any of these assumptions. No effort has yet been made to develop a broad national consensus among faculty regarding definitions of critical thinking and communicating, much less about ways to teach these concepts. Moreover, current measurement theory and its application in the development of instruments designed to assess postsecondary students' general intellectual skills are inadequate to provide specific direction for improving either teaching or learning. Certainly the act of assessing student abilities will not, in and of itself, improve those abilities.

Nevertheless, if decision-makers determine that the national interest will be served by a comprehensive postsecondary assessment program, the principles of continuous improvement heretofore applied with most success in industry should be used to link faculty goal-setting, staff development related to teaching, assessment of instructional resources and processes as well as student outcomes, and use of assessment results to improve teaching and learning. This paper provides a rough sketch of a national assessment-and-improvement program based on these principles that might be developed over the next several years.

# TOWARD A PLAN FOR USING NATIONAL ASSESSMENT TO ENSURE CONTINUOUS IMPROVEMENT OF HIGHER EDUCATION

Trudy W. Banta

## The Problem of Assessing College Students' Abilities

Embedded in the fifth objective of National Education Goal 5, which is stated, (By the year 2000) "the proportion of college graduates who demonstrate an advanced ability to think critically, communicate effectively, and solve problems will increase substantially," are at least five implicit assumptions. This paper begins by examining these assumptions in some detail and raising questions about the viability of each under prevailing conditions in American higher education. In a second part of the paper, recommendations are made for addressing these concerns in a national assessment-and-improvement project designed to promote the achievement of Objective 5.5.

The assumptions implied in Objective 5.5 of the National Goals that will be considered here include the following:

1. The abilities to "think critically", "communicate effectively", and "solve problems" can be operationally defined and these definitions agreed upon as desired instructional outcomes by all U.S. faculty responsible for developing these abilities in undergraduates.

2. The defined abilities will be taught, by all faculty charged with the responsibility for teaching them, in ways that engage students and promote learning of these abilities.

3. Reliable and valid measures of student achievement of the defined abilities can be identified or created.

4. The measures of student attainment can be administered to all college graduates (or samples of that population) in settings that engage students and encourage their best efforts.

5. The results of assessment of developed student abilities will be used to improve the materials and methods of instruction in ways that increase student engagement and promote learning gains.

The discussion of these points which follows is limited by the author's current perspective. Since 1982, I have coordinated a comprehensive student outcomes assessment program at the University of Tennessee, Knoxville (UTK). In terms of its longevity, the extent of participation by units within the institution, numbers of students tested, and comprehensiveness of its on-going assessment-related research agenda, the outcomes assessment program at UTK is unique among those at U.S. research universities. Over the years, more students have been tested with a broader array of standardized measures of general intellectual skills--including dimensions of critical thinking, communicating, and problem-solving abilities--at UTK than at any other institution in the country. The primary impetus for this extraordinary institutional investment in assessment is the Tennessee Higher Education Commission's performance funding program, which began in 1979 to provide the basis for an annual supplement to the budget of each of the state's public colleges and universities for conducting specified outcomes assessment activities (Banta, 1988). In 1991, the budget supplement available to UTK through the performance funding program was approximately $6 million.

In the ensuing sections, many of the nuances of the debate surrounding implementation of strategies to advance Objective 5.5 are omitted, or at best treated superficially. Other writers on the panel have the expertise to illuminate those areas. My contribution reflects most directly my own experiences in several areas that will be of critical importance in implementing Objective 5.5. That experience includes 1) working

with faculty to select or design measures of the general intellectual skills of college graduates, 2) working with students to increase their levels of motivation to do their best work on a required comprehensive exam that is not an integral part of their coursework, 3) working with a staff of test administrators to ensure that students encounter testing conditions that are conducive to their best performances and 4) designing and administering a research program aimed at improving practice in postsecondary outcomes assessment. Given my deep personal commitment to the improvement of practice, the emphasis throughout this paper is on the process--what evidence do we have that the current process of assessing postsecondary student outcomes is capable of producing improvements in those outcomes; and if that process is inadequate, how might positive change be effected?

Each of the five assumptions outlined previously is treated in a section of the first part of this paper. Evidence from the literature and/or practice is cited to support the concerns raised in connection with each assumption.

## The Abilities Can Be Defined and Agreed Upon

To correct a process, one must be able to define it clearly and identify its elements. The definition and elements then must become agreed-upon goals and objectives for behavior and action. Systematic development of and adherence to explicit goals for courses and curricula are not currently pervasive practices in higher education. Most faculty are not trained specifically for the job of teaching--though more graduate programs are now providing such training--and many simply are not aware of the importance of setting

specific goals and objectives as a first step in preparing a course or curriculum outline (Boyer, 1990).

Faculty who do develop goals and objectives usually do not state them in terms of what students should know and be able to do as a result of experiencing the course or curriculum (Gardiner, 1989). The goals/objectives are much more likely to be statements of process that say what the instructor will do--what content will be presented in class, what assignments students will be given. Finally, statements of goals and objectives are not always shared with students; thus the students are not aware of the precise nature of what they are expected to learn. Students cannot be purposeful about their learning in the absence of purpose statements provided by faculty.

For more than ten years, faculty in Tennessee public colleges and universities have been aware that the performance funding program administered by the Tennessee Higher Education Commission requires them to prepare students for senior exams in general education and their major field. If any faculty has had the time and encouragement to approach assessment of student outcomes systematically, it is the one at the University of Tennessee, Knoxville. Yet a recent survey of department heads on that campus has revealed that no more than 30 percent of the faculty have developed explicit written student outcome objectives for their courses or curricula (Center for Assessment Research and Development, 1991).

Even when faculty have developed expertise in the techniques of stating objectives in the form of student outcomes, each instructor usually prepares his/her own goals in isolation. The tradition of cooperation on such matters is almost non-existent in higher

education. A national survey of faculty conducted in 1989 for the Carnegie Foundation (Boyer, 1990) revealed that 44 percent of the respondents agreed that "Faculty in my department have fundamental differences about the nature of the discipline."

The possibility of gaining a national consensus on stated goals/objectives for promoting critical thinking seems virtually impossible. Cuban (1984, p. 676) has called this area a "conceptual swamp." The experience of the American Philosophical Association is telling in this connection: after a six-round Delphi process that took place over a period of nearly two years, 46 professionals with expertise in critical thinking instruction, assessment, or theory were able to develop a "Consensus Statement Regarding Critical Thinking and the Ideal Critical Thinker," but the principal investigator, Facione (1990), revealed that even where consensus was reported, a minority of panelists held divergent views.

It is worth emphasizing at this point that while the difficulty of achieving a national consensus on the meaning of such a complex ability as critical thinking is enormous, building that consensus is absolutely essential. If significant improvements at the national level are desired, all faculty responsible for developing the abilities specified in Objective 5.5 must subscribe to the national goals and objectives. Indeed every student deserves an equal opportunity to experience the curricula and courses designed to promote the achievement of these goals, and no less than a nation-wide effort will be needed to bring about the "substantial" increases in the abilities that are desired by 2000.

I leave to others the fuller discussion of the problems of reaching consensus on definitions of Objective 5.5 abilities. I do agree with Cuban (1984) and others, however,

that critical thinking, reasoning, and problem-solving are virtually indistinguishable. Therefore, throughout the remainder of this paper, I shall use the term critical thinking to refer to both the critical thinking and the problem-solving components of 5.5. In the section of the paper that discusses specific postsecondary outcomes assessment instruments, the terms critical thinking and communicating refer perforce to the operational definitions given to those concepts by the developers of the tests reviewed there.

**The Defined Abilities Will Be Taught**

If it were possible to develop a national consensus among faculty concerning the definitions of the abilities of critical thinking and communicating, the faculty in any given college or university would need to identify those courses and course sequences at their institution that should promote student learning in those areas. Alverno and King's College faculties have defined generic abilities and have designated those points in the curriculum at which students will experience each (Alverno College Faculty, 1979; Farmer, 1988). Thus there is some evidence that it is possible for faculty at a given institution to agree upon what should be taught, by whom, and when. But at larger and more complex institutions, especially research universities, faculty are not likely even to recognize the need for such agreement, must less to come together to establish it.

Faculty prize their autonomy. Many at comprehensive universities work alone on their research, or perhaps with colleagues in their discipline at other institutions. In the name of academic freedom, they maintain their right to pursue their own lines of inquiry, both in their scholarship and in the courses they teach. Some would even use this

argument to oppose vertical integration of the curriculum in a given area, that is, the extent to which lower-division courses in a sequence are linked with their upper-division counterparts and thus provide students with specific experiences that prepare them for more advanced courses.

Not only are many faculty reluctant to have others suggest what they should teach, they also have limited interest in the formal pursuit of learning how to teach. Having completed graduate programs in which individual scholarship was the principal focus, few have spent very much time thinking about or studying how college students learn and how teaching can promote their learning. Eble (1972, p. 180) decried the "narrowness of vision, the disdain for education, the reluctance to function as a teacher" that he considered "ills attributable in large part to graduate training." In all fairness, 58 percent of the faculty at 4-year institutions who responded to the 1989 Carnegie Foundation survey said their primary interest was in teaching as opposed to research, which most interested 42 percent (Boyer, 1990). But institutional reward systems make it difficult for faculty to spend as much time as they might wish to spend improving their effectiveness as instructors because tenure and promotion criteria emphasize research and scholarly attainment at the expense of teaching.

To complicate matters, today's students have grown up with a steady diet of fast-paced video-based news and entertainment programming, and thus are increasingly bored by lectures--the preferred presentation format of the traditionally-prepared professoriate. Add to this the disaffection with the academic environment created by the part- or full-time work in which so many students engage, and one can begin to sense the scope of the

problem in motivating students to become engaged in their learning. This pervasive lack of motivation among American students is an abiding concern, even of the most thoroughly prepared and student-oriented instructors.

**The Abilities Can Be Measured**

Higher education decision-makers are interested in assessing the general intellectual skills of substantial numbers of college students reasonably quickly and at modest cost. They would like to obtain scores that are easy to interpret and comparable across individuals and groups. The scores should be reliable and valid--measuring what they purport to measure--and suggest directions for action aimed at improving students' scores. Ideally, there should be ways to compare scores for individuals over time to assess their progress as a result of their educational experiences.

Few of these desired characteristics are attained in the measures of general intellectual skills that are currently used in postsecondary outcomes assessment programs.

Since the late 1970s, when assessment of student outcomes began to emerge as an important component of the accountability movement in American higher education, four standardized tests have been developed and marketed in response to the need of institutions for instruments that assess the general intellectual skills of substantial numbers of students reasonably quickly and at modest cost. Standardized exams, as opposed to tests that faculty might develop locally, also offer scores that are relatively easy to interpret and norms that permit individual, program, and institutional comparisons. The four instruments that have come to be used most widely in postsecondary outcomes assessment

are the College Outcome Measures Program (COMP) exam and the Collegiate Assessment of Academic Proficiency (CAAP), developed by the American College Testing Program (ACT); the Academic Profile, developed by the Educational Testing Service (ETS); and the College Basic Academic Subjects Exam (CBASE), developed at the University of Missouri-Columbia and marketed by Riverside Publishing Company. Since 1988, all of these tests have been administered to seniors at the University of Tennessee, Knoxville and have been systematically evaluated by faculty and students (Banta and Pike, 1989).

UTK has an undergraduate student population of about 19,000, and approximately 3000 of these students graduate each year. Since 1985, all seniors have been required to take a test in general education prior to graduation. Scores are reported annually to the Tennessee Higher Education Commission, and the level of the scores is the factor that determines the proportion UTK will receive of a performance funding budget supplement of more than $1 million from the state. Testing takes place on Saturdays and weekday evenings, on campus but outside the framework of academic coursework. Students receive in freshman orientation and in advising sessions information about the University's emphasis on outcomes assessment and the importance of the senior exam in that program. The registrar notifies each senior by mail of the need to take the test in general education prior to graduation. Though some UTK seniors are still completing general education coursework, they are encouraged to take the exam as early as possible in their senior year in order to avoid scheduling conflicts during their last term.

Since the program of pilot-testing multiple instruments and comparing their technical qualities began at UTK in 1988, seniors have been randomly assigned at the time

of testing to take one or the other of the two exams under study in a given year. Annually, some 100 volunteers have received gift certificates for taking both exams; this double-testing has permitted the calculation of inter-scale correlations for the two exams. All seniors have been asked to respond to a series of written questions concerning their assessments of the content of the tests. In addition, faculty with special interest in the University's general education curriculum have evaluated the content representativeness of the exams, comparing the content of each with the institution's statement of goals for general education (Banta & Pike, 1989).

All four of the standardized exams under scrutiny purport to measure some dimensions of the critical thinking and communication skills of college students. Scales on the COMP exam include Functioning in Social Institutions, Using Science and Technology, Using the Arts, Communicating, Solving Problems, and Clarifying Values. The CAAP includes Writing, Mathematics, Reading, and Critical Thinking scales. The Academic Profile is composed of scales labeled College-Level Reading, College-Level Writing, Critical Thinking Using Mathematical Data, Humanities, Social Sciences, and Natural Sciences. Scales that make up the CBASE are English, Mathematics, Social Studies, Science, and interpretive, strategic and adaptive reasoning.

Reliability estimates for the total scores derived from these four exams are acceptable, ranging from .84 for the COMP to .94 for the Academic Profile (Pike, 1991a), though they are somewhat lower than those associated with such established measures as the ACT, SAT, and Graduate Record Exam. However, reliability of subscales is lower than for total scores, and in some cases the level is unacceptable, as with the .44 for the COMP

subscale Clarifying Values, which is a component of the COMP approach to measuring critical thinking (Pike, 1989). Moreover, factor analysis reveals that only the CBASE is composed of subscales that actually measure what they purport to measure. The other tests assess a single factor, which seems most likely to be verbal ability, and even CBASE scores are highly correlated with students' entering levels of ability (Pike, 1989, 1990, 1991a). No more than one-fourth of the UTK seniors taking any of the general intellectual skills tests considered it a good or excellent measure of their knowledge and skills in general education, and faculty concluded that none of the tests assessed more than one-third of the content specified for inclusion in the University's general education program.

Over the past thirty years, measurement theorists have spent considerable amounts of time and energy debating the issue of whether skill in critical thinking is more dependent upon deep expertise in a specialized area or upon possession of well-developed generic reasoning strategies (Perkins and Salomon, 1989). The theoretical debate has been extended to include applications in teaching and assessment methods. The COMP exam is a test of "effective adult functioning," and employs items that are less content-specific than those used in the CAAP, the Academic Profile, and the CBASE. Thus generic as well as domain-specific approaches to measuring critical thinking are represented in these four exams.

Regardless of the measurement approach utilized, however, our studies show that students' scores on all four tests are much more highly related to initial ability than to any other factor. Attempts to trace the impact on these scores of coursework and other educational experiences associated with the college years have not yielded definitive

answers (Pike, 1991b). Hanson (1988) attributes this failure to the fact that today's test developers know best how to measure static traits, such as verbal ability, as opposed to developmental changes. Since measures of static traits are based on the assumption that the underlying structure of the construct being measured does not change over time, such measures may not be able to detect student characteristics that change as a result of college experiences.

The evidence assembled to date from research and experience in postsecondary outcomes assessment leads to the conclusion that current measurement theory and its application in the development of instruments designed to assess students' general intellectual skills are inadequate to support specific suggestions for improving students' learning based on their scores on these instruments.

**Measures of the Defined Abilities Will be Taken Seriously by Students**

Another problem encountered in attempting to administer standardized tests intended to serve the purpose of outcomes assessment to groups of college students is that in the absence of explicit connections between their performance on the test and their academic program, students see little need to do their best work (Warren, 1989). Providing money or other extrinsic rewards as incentives may initially motivate some students, but the novelty wears off quickly.

When college seniors are required to take a standardized test for purposes of evaluating their general education program, and their performance on that test has no consequences in terms of their progress in a course or program, ten years of experience at

UTK suggests that only one-fourth may be willing to try their hardest (Banta and Pike, 1989). What credence can be given to scores derived from a population of test-takers, the majority of whom are at least indifferent to the need to apply themselves to the task, if not determined to deliberately falsify their responses?

**Assessment Will Increase Student Learning**

Current measurement theory and technology do not support a value-added approach to assessing learning gains over time, either for individual students or for groups of students (Lord, 1967; Cronbach and Furby, 1970; Warren, 1984; Baird, 1988). Cross-sectional studies suffer from the inability of research designs to account for all the differences between cohorts that may influence test performance, and even longitudinal studies that examine change in the same individuals between two points in time are plagued by serious technical problems. A partial listing of these problems includes Warren's (1984) concerns that students may not have a sufficient knowledge base against which change can be measured, and that when significant differences in knowledge do exist, the scores of students at opposite ends of the knowledge continuum cannot be compared because they are qualitatively different. The spurious negative correlation between initial status and score gain that obscures the meaning of gain in studies of student growth is one manifestation of Warren's second concern.

Only equivalent forms of the same test can be used to provide clear evidence of student growth due to the effects of education programs, and even for the carefully-developed National Assessment of Educational Progress (NAEP), it has not been possible to construct forms that are truly equivalent (Zwick, 1991).

In a recent study of longitudinal change scores on the COMP exam at the University of Tennessee, Knoxville, Pike (in press) applied the three most widely used techniques for assessing student growth and development--gain scores, residual scores, and repeated measures--and found serious weaknesses in each.

If it is not possible to use today's standardized tests to document specific changes in student learning that take place as a result of educational experiences, then there is little if any basis for using these scores to suggest improvements for instructional methods or materials. Certainly the act of assessing student learning will not, in and of itself, improve that learning. A decade of sporadic, unconnected assessment activities in higher education and at least two decades of achievement testing in grades K-12 serve well to illustrate this point. In fact, there is growing concern that the vast network of testing programs in elementary and secondary schools in this country has actually been an influential factor in lowering academic standards to the level of what can be easily and reliably assessed (Moss and Koziol, 1991; Nickerson, 1989), with a consequent overall negative impact on teaching and learning (Frederiksen and Collins, 1989).

## An Approach to the Task of Assessing College Students' Abilities

The problems involved in developing appropriate measures of critical thinking and communication skills for college graduates have been identified, and they are daunting. Some of the most knowledgeable measurement specialists say that it is not currently possible to develop an assessment program that meets the twin goals of monitoring status for accountability purposes and providing direction for instructional improvement because

optimizing validity for one purpose diminishes it for the other (Moss and Koziol, 1991). Hanson (1988, p. 54) believes that

> assessing when and how students change, and linking such change to specific educational interventions, is a complex and difficult task that requires new strategies for conceptualizing issues, building new and different assessment instruments, and designing research with different purposes and outcomes than those found in many traditional methods of inquiry.

The possibility of solving the measurement problems associated with using testing to improve learning should not be rejected just because they are so difficult. Moreover, the nation's governors want measures of college student learning to be developed. The President and the Secretary of Education have implied that such measures will be developed in their formal statement of National Goals for 2000. And apparently three-fourths of the American people believe that nationally standardized tests for students can play an important role in improving education in this country (Elam, Rose, & Gallup, 1991).

Monitoring progress, or assessing status, is a component of any effective process. But if there is anything that Edwards Deming (1986), Japanese industrialists, and winners of the Malcolm Baldrige Award in this country have taught us in recent years, it is that inspection alone will not produce improvement.

If we are going to make the investment to create a national monitoring system focused upon the higher-order cognitive skills of college graduates, then we must secure that investment by making the monitoring activity part of a larger system that ensures the use of assessment findings to improve education. That is, as implied in the first part of this paper, we must specify clear goals and objectives for the skills we seek in college

graduates, we must provide the staff development and instructional resources necessary to prepare faculty to teach these skills using methods that genuinely help students learn them, we must develop precise measures of the specified skills and administer these to students in ways that encourage their best efforts, and then we must use the results of assessment to modify the components of this system that are shown to be in need of improvement. This will require a national effort of epic proportions. It will be enormously costly. But if we are determined to attack this problem, and to do so in ways that have a chance of being effective, we must begin systematically, drawing upon everything that recent experience with assessment at elementary, secondary, and postsecondary levels has taught us.

The sections that follow suggest in very rough outline some strategies for a comprehensive assessment-and-improvement program designed to secure the investment in a national postsecondary monitoring activity. The suggestions fly in the face of current tradition and practice in higher education. Nothing less than a cultural change will be required to carry them out successfully. However, the time may be right to effect such a change.

## Setting the Goals - Describing the Well-Prepared Graduate

Secretary Lamar Alexander, with his proven ability to capture national attention for ambitious goals and programs, and David Kearns, with his experience in establishing continuous improvement of quality as an organizational philosophy at Xerox, bring a unique combination of leadership skills to the task of mobilizing the higher education

community for the work of implementing Objective 5.5. They also hold the club that can be brandished if that community fails to respond: the threat of withdrawal of federal funds.

A federal "5.5 Panel" should be appointed, with representation from the groups most concerned about the preparation of college graduates. Examples of these groups include students themselves, parents, employers, faculty, and K-12 educators. Each governor should also appoint a state level panel similarly constituted.

Drawing upon their own experience and previous efforts to define a domain of knowledge ( Adelman, 1989; Tennessee Higher Education Commission, (1977); Facione, 1990; Alverno College Faculty, 1979; Farmer, 1988; Peterson, 1982), the national and state 5.5 Panels should first describe critical thinking/problem solving and communicating in terms of what a competent adult should know and be able to do in each of these areas. Employers and parents, as just one example of a pair of constituent groups with diverse perspectives, may start by describing knowledge and behaviors in different terms, but ultimately it should be possible to reach some consensus.

Next, the 5.5 panelists should ask themselves, "How will we know, how can we be satisfied, that a young adult possesses the knowledge and exhibits the behaviors we have specified?"

Substantial national involvement in defining the critical abilities and suggesting how their attainment might be assessed can be achieved if the state 5.5 Panels solicit ideas, then reviews, of preliminary work from the public, but especially from faculty at public and private colleges. A Delphi process may be helpful in this endeavor (Facione, 1990). Staff

of the federal panel can synthesize the work of the 50 state panels and the federal panel, and a final review and approval process can be specified by the federal group. Besides helping to develop a feeling of ownership for the national goals on the part of local faculty--an essential element of this effort since teachers will not teach what they do not value (Wiggins, 1990)--creating the state panels would offer the advantage of providing a wide variety of suggestions for ways to measure achievement of the goals. A great deal of imaginative effort will be needed in this area.

**Preparing Faculty to Foster Student Learning**

The quality improvement literature emphasizes that since people generally want to do their jobs as well as they can, most of the obstacles to fulfilling this ambition are not the fault of the people involved but rather of the systems in which they must work (Deming, 1986; Imai, 1986). The individuals employed as faculty in our colleges and universities have been socialized in an academic tradition that rewards individual achievement and intellectual and behavioral autonomy (Eble, 1972; Boyer, 1990). Substantial incentives must be provided if faculty are to work together on plans to implement strategies designed to foster the development of specified critical thinking and communication abilities. Adoption of Boyer's (1990) proposal that reward structures in higher education be modified to include more emphasis on the scholarship of teaching would be very helpful in this connection.

The best instructional development specialists and the most outstanding post-secondary teachers in the country should be assembled in Washington to develop strategies

for teaching the knowledge and behavior described in the final report of the federal 5.5 Panel. Considerable guidance for this work is available in such contemporary sources as Brown, Collins, and Duguid (1989), Miller and Gildea (1987), Perkins and Salomon (1988), and Sternberg (1985a and 1985b).

Given the knowledge and skill definitions in the 5.5 Panel Report, every college and university faculty should decide upon its own program of in- and out-of-class experiences that will promote student development of the specified abilities. Selected faculty and staff should be charged specifically with the responsibility for providing these experiences in courses and out-of-class activities. State teams of staff development specialists should be trained for the task of acquainting college faculty in a given state with the teaching strategies and materials developed by the group assembled at the federal level. If special facilities, equipment, or materials are deemed essential in enhancing teaching and/or learning, these should be provided on every campus.

Continuous student and faculty review and evaluation of teaching strategies and materials must be built into this process. And as experience proves certain approaches to be more valuable than others, this information should be used to modify the curriculum used by the state staff development specialists.

## Gathering Evidence of the Process and Outcomes of Student Learning

The process of student learning. A point emphasized throughout this paper is that inspection of student attainment at the end of the educational experience provides woefully inadequate information for directing improvement efforts. While culminating assessment

activities must be developed and administered in accomplishing the intent of Goal 5.5, the goal cannot be fully realized if additional data about the context for student learning are not collected. Grandy (1989) has argued that assessment must be closely linked with specific elements of student learning if causal connections that suggest directions for remedial learning strategies are to be made evident. Warren (1989, p. 65) believes that "What is taught, how intensively, for what length of time, in what way, using what resources" are all essential influences on student learning that must be assessed if we hope to assemble sufficient data to stimulate improvements in the educational process.

A federal panel of outstanding measurement specialists and college and university faculty should be assembled to map the program of assessment strategies that will be necessary to realize Objective 5.5. In keeping with the foregoing suggestions from the literature, this 5.5 Measurement Panel should set up a reporting system to gather institution-specific responses to the following questions:

1) Is student growth a clearly articulated and implemented institutional goal?

2) Is each student and faculty member aware of the federal expectations with respect to student development of Objective 5.5 abilities?

3) How much time has each faculty member spent in staff development activities related specifically to promoting students' learning of Objective 5.5 abilities?

4) How much time does each faculty member spend preparing to teach, and teaching, material related to Objective 5.5?

5) How much time does each student spend studying material related to Objective 5.5?

6) How much out-of-class time does each student spend in conversation and/or activities related to the 5.5 abilities?

7) Do students and faculty perceive that they have access to the facilities, equipment, experiences and materials they need to promote development of 5.5 abilities?

8) Is student progress toward development of 5.5 abilities sufficiently evaluated, and is the student briefed concerning that progress?

9) Are students sufficiently motivated to develop the 5.5 abilities and to do their best work when their progress is evaluated?

Just as individual students must assume responsibility for developing the 5.5 abilities if they expect to graduate from college, faculty and staff associated with individual campuses, and programs on those campuses, must take responsibility for gathering the data that will enable them to understand what actions they can and should take to maximize student growth and development in these areas. A variety of data sources will be needed to provide answers to questions 1-9 above; examples of some of these are given below.

Peterson and Cameron's (1988) "Organizational Climate for Teaching and Learning" and the "Inventories of Good Practice in Undergraduate Education" associated with the Wingspread "Seven Principles for Good Practice in Undergraduate Education" (Chickering, Gamson, & Barsi, 1989) provide examples of the kinds of questions that might be asked of faculty, staff, and students to ascertain an institution's commitment to student growth (see Question #1 above). Question 2 can be answered by asking students and faculty to summarize the federal expectations as they understand them.

State and local staff development specialists will have records that show the amount of formal training each faculty member has experienced in connection with learning how best to foster student learning of the 5.5 abilities (#3 above). Faculty members themselves must supply a total number of such hours spent, however, because they may have engaged

in additional formal or informal developmental activities beyond those provided by the state or their institution. Such additional activities should be described, because they could prove to be more effective than the state-initiated programming.

Question 4 requires data from several sources. Faculty with responsibility for developing 5.5 abilities can report the number of hours spent preparing to teach and teaching material related to these abilities. Course syllabi can be examined to ascertain the relative emphasis given the development of 5.5 abilities as compared with the attention given to other topics. Finally, students can be asked to record the amount of time they spend studying material related to 5.5 abilities both in class and outside class (#5 above).

Student involvement in learning 5.5-related knowledge and skills (#6 above) can be gauged via items similar to those in the College Student Experiences Questionnaire (Pace, 1990). Question 7 can best be answered by asking students and faculty directly about the adequacy of facilities, equipment, in - and out-of-class experiences, and materials.

Examinations and student assignments in courses designated to make contributions to student development of 5.5 abilities should be reviewed to ascertain that they contain appropriate evaluations of student progress (#8 above). Moreover, the nature and extent of information about progress that is given to students should be described, both by the initiating faculty and by the student recipients.

Finally, students should be asked specifically about their level of motivation to do their best work generally, and with respect to developing and exhibiting 5.5 abilities specifically (#9 above). Even the most carefully-constructed sequence of learning activities will not promote the development of desired abilities in students who are not motivated to benefit from the activities (Warren, 1989).

The outcomes of student learning. If implemented, national assessment designed to accomplish Objective 5.5 will constitute high-stakes testing for accountability purposes for the nation's colleges and universities. In grades K-12, this kind of testing has been shown to influence teaching behavior (Nickerson, 1989). Wiggins (1989) argues that if tests are going to determine what teachers teach and what students study, the tests should focus on capabilities and habits that we consider essential for students to master. Fredericksen and Collins (1989) have written of systemic validity as a test property which indicates that an instrument induces curricular and instructional changes that promote development of the cognitive skills the test is designed to measure. These authors contend that indirect, objectively-scored tests excessively narrow what is taught and learned, and direct measures, subjectively scored, maximize systemic validity.

The work of all the investigators just mentioned, plus that of Brown, Collins, and Duguid (1987), Miller and Gildea (1987), and many others supports the development of a new examination system that emphasizes alternatives to traditional multiple-choice instruments. At the institutional level, colleges and universities should use course-embedded assessment to monitor the process of student development of 5.5-related skills. In addition, for national accountability purposes, every graduate should complete a written thesis or project as a capstone experience during the final year in college. This project might be supplemented by a narrative or videotaped portfolio (Learning Research and Development Center, 1990) that would supply a window on student development over time, thus contributing the additional dimension of a value-added approach.

At Alverno (1979) and King's (Farmer, 1988) Colleges, embedding assessment activities in coursework has proven to be the most compelling means of ensuring that faculty will teach and students will learn fundamental abilities such as critical thinking and communicating. At these institutions, instruction in generic skills is competency-based, students are informed about the abilities they are expected to develop, and assessment takes place at appropriate points in their courses. The level of motivation to do well on assessment activities is high because students understand that these activities are important--faculty have given them value by including them in assignments and tests that count in course evaluations.

In implementing a national project aimed at promoting achievement of Objective 5.5, undergraduate students must be apprised early in their academic careers of the precise definitions of the abilities they are expected to develop by the time they graduate. Experiences designed to promote these abilities should be explicit inclusions in early courses and out-of-class activities. Students should receive the instructions and scoring criteria for the senior/thesis project well in advance of the senior year. They should know that they themselves are responsible for developing the skills and knowledge implicit in the scoring criteria and that their success in doing so, as demonstrated in their performance on the senior project, will have a significant bearing on their attainment of the status of college graduate.

The federal 5.5 Measurement Panel should develop specifications for assigning and scoring the senior project based on parameters established in the federal 5.5 Panel Report. Instructions for preparing the senior project should elicit from the student expressions of

all of the essential elements of critical thinking and communicating described in the 5.5 Panel Report. In addition, the Measurement Panel must carefully and explicitly define acceptable (and unacceptable) levels of performance for each of the abilities. The work of Fredericksen (1986), Warren (1984), and developers of NAEP scoring methods (Braun, 1986; Breland & Jones, 1988) should prove instructive in this endeavor (though Forsyth's (1991) reservations about the NAEP test development and scoring processes should certainly be noted).

The senior thesis/project should be evaluated by at least two trained readers at the institution--preferably individuals not involved in teaching the senior course in which the assignment was given. In order to promote individual learning, each student should receive a detailed review of his/her performance on the project (Alverno College Faculty, 1979; Stone & Meyer, 1989). The criteria used to evaluate student work should, of course, be those developed by the 5.5 Measurement Panel.

The detailed reviews of student projects should be read next by an institution-wide committee charged with the responsibility of identifying strengths and weaknesses in the preparation of graduates generally, and within each major where numbers warrant. Warren (1989) has described benefits associated with categorical grading--reading one item at a time across students and/or classes--which is a procedure that could be applied here.

A state-wide committee should read a randomly-selected sample stratified by major of students' papers from each institution in the state. Institutions should be informed of the relative performance of their seniors on each of the specified criteria as compared with performance at other institutions in the state.

After substantial work on establishing inter-rater agreement among state and federal reviewers, the latter being members of the 5.5 Measurement Panel who have established the evaluative criteria, the performance ratings of each state could be compiled to yield a national composite for each of the specified abilities. Moss and Koziol (1991) have described methods for increasing inter-rater agreement on such a task, but have also noted the difficulties involved in comparing student performance across different tasks; the relationship among scores within an essay exam is stronger than the relationship between essay exams.

**Making Assessment Count: Using the Findings to Effect Improvements**

The Secretary of Education must keep the sights of the postsecondary community focused on continuous improvement of student performance on the 5.5 abilities. The Measurement Panel will set the criteria for acceptable national performance, and if these are met, the standards should be raised, or new criteria should be formulated. The purpose of this national assessment effort cannot be simply to report on status. The initial report on student performance must mark the beginning of a significant program of focused educational improvement.

Experience over the past decade with assessment at the postsecondary level has indicated that the findings or results obtained from assessment are less important in stimulating improvements in practice than is the process of bringing faculty together to discuss purposes, student outcomes, and methods of instruction as they prepare for outcomes assessment (Banta and Fisher, 1986). Nevertheless, connections can be made

on each campus between the independent process variables outlined in the preceding section and the dependent performance variables derived from analysis of the senior thesis/project and from the accompanying portfolio materials if these are included. Cluster-analytic methods (Ratcliff and associates, 1988)and hierarchical linear models (Raudenbush & Bryk, 1989) can be used at the individual campus (or program) level to identify combinations of factors and experiences that promote student development of specified knowledge and skills. This national assessment project, if undertaken as outlined, could produce an unprecedented opportunity to identify factors and experiences that enhance teaching and learning.

The Secretary of Education should make available to each state a significant amount of money each year to support improvement efforts proposed by colleges and universities that have analyzed their assessment data thoroughly and thus can provide evidence that their proposals are likely to result in increased student learning. If staff development is revealed to be a significant need for many campuses, as is anticipated, the Department of Education should consider supporting the development of improved national resources in this area. Frederiksen and Collins (1989) have advocated using the materials developed originally for the purpose of training faculty to employ specific standards in assessing student work in a secondary capacity to provide professional development experiences designed to help faculty strengthen their own teaching and classroom assessment strategies.

### Assessing the Non-College-Going Cohort

In order to calculate an estimate of the effects of college, as opposed to those of

simple maturation and life experiences, on the development of 5.5 abilities (Pascarella and Terenzini, 1991), a measure of these abilities needs to be derived from the non-college-going peers of college graduates. The national assessment effort just described will take several years to develop in colleges and universities. As funds permit, the same senior thesis/project given to college seniors should be assigned to members of the age cohort (probably 24-year-olds if this is determined to be the average age of college graduates) who are employed in the military and in the nation's largest companies. A criterion for identifying a large company might be that it employs 50 or more 24-year-olds.

Employed 24-year-olds should receive the same advance notice of the need to complete the thesis/project as college seniors receive, with the attendant instructions and specifications. Employers who see the national assessment project as a long-range strategy for improving the preparation of the college-educated workforce should be able to provide the motivation for their 24-year-old employees to take the senior project as seriously as do college seniors.

The 5.5 Measurement Panel would need to supply professionals thoroughly trained to use its scoring criteria to serve as readers of the projects completed by the non-college-going age cohort. Every individual completing a project should receive detailed information about his or her own performance. Units of organizations having 50 or more 24-year-olds should receive their aggregated scores on each of the specified components of the 5.5 abilities.

## A Concluding Observation

The national postsecondary assessment strategy proposed here is a multi-billion-dollar undertaking. Since much of the investment would be in people's time at the institutional and state levels, the cost at the federal level would represent only a minuscule portion of the total investment. Initially there must be broadly-based consideration of the advantages and disadvantages associated with having so many faculty members substitute the staff development, planning, data-gathering, and analysis activities of this project for other activities in which they are presently engaged. Satisfying the national interest in imposing the development of a core of common competences upon all college graduates must be balanced against the potential loss of academic freedom for individual faculty (Miller, 1991) and the possible reduction in diversity among institutions with a wide variety of missions that have been the hallmarks of this country's widely-admired system of higher education.

If the decision is made to invest the national resources necessary to carry out the comprehensive approach to assessment-and-improvement proposed in this paper, the Secretary of Education would have an opportunity to establish the boldest and potentially most promising research and development project ever undertaken in higher education. This work could establish the basis for making continuous improvement a part of everything that is done in the name of postsecondary education. This development could help colleges and universities reclaim some of the responsibilities for providing higher education that they are losing to private industry and federal agencies. Finally, this approach could ensure that the higher education system in the United States will be sufficiently responsive to changing global needs to maintain its current reputation as the best in the world.

## References

Adelman, C. (1989). Introduction: Indicators and their discontents. In C. Adelman (Ed.), Signs and traces: Model indicators of college student learning in the disciplines, (pp. 1-10). Washington, D.C.: U.S. Government Printing Office.

Alverno College Faculty. (1979). Assessment at Alverno College. Milwaukee, WI: Alverno Productions.

Baird, L. L. (1988). Value-added: Using student gains as yardsticks of learning. In C. Adelman (Ed.), Performance and judgment: Essays on principles and practice in the assessment of college student learning, (pp. 205-216). Washington, D.C.: U.S. Government Printing Office.

Banta, T. W. (1988). Assessment as an instrument of state funding policy. In T.W. Banta (Ed.) Implementing outcomes assessment: Promise and perils. New Directions for Institutional Research, 59. San Francisco: Jossey-Bass.

Banta, T. W. & Fisher, H. S. (1986). Assessing outcomes: The real value-added is in the process. Proceedings from the Conference on Legislative Action and Assessment: Reason and Reality (pp 81-91). Arlington, VA: The American Association of State Colleges and Universities.

Banta, T. W. & Pike, G. R. (1989). Methods for comparing outcomes assessment instruments. Research in Higher Education, 30(5), 455-469.

Boyer, E. L. (1990). Scholarship reconsidered: Priorities of the professoriate. New Jersey: Princeton University Press.

Braun, H. (1986). Calibration of essay readers (Report No. RR-86-9). Princeton, NJ: Educational Testing Service.

Breland, H. M. & Jones, R. J. (1988). Remote scoring of essays (Report No. 88-4). Princeton, NJ: Educational Testing Service.

Brown, J. S., Collins, A. & Duguid, P. (1989). Situated cognition and the culture of learning. Educational Researcher, 18(1), 32-42.

Center for Assessment Research and Development. (1991). Assessment of institutional effectiveness at the University of Tennessee, Knoxville. In SACS self-study. Knoxville: The University of Tennessee.

Chickering, A. W., Gamson, Z. F. & Barsi, L. M. (1989). Inventories of good practice in undergraduate education. Racine, WI: The Johnson Foundation.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change" - or should we? Psychological Bulletin, 74(1), 68-80.

Cuban, L. (1984). Policy and research dilemmas in the teaching of reasoning: Unplanned designs. Review of Educational Research, 54, 655-681.

Deming E. W. (1986). Out of the crisis. Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Eble, K. E. (1972). Professors as teachers. San Francisco: Jossey-Bass.

Elam, S. M., Rose, L. C., & Gallup, A. M. (1991). The 23rd annual Gallup poll of the public's attitudes toward the public schools. Phi Delta Kappan, 73,(1), 41-56.

Facione, P. A. (1990). Executive summary of "Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction." Millbrae, CA: California Academic Press.

Farmer, D. W. (1988). Enhancing student learning: Emphasizing essential competencies in academic programs. Wilkes-Barre, PA: King's College.

Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? Educational Measurement: Issues and Practice, 10(3), 3-9,16.

Frederiksen, J. R. & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18(9), 27-32.

Frederiksen, N. (1986). Toward a broader conception of human intelligence. American Psychologist, 41, 445-452.

Gardiner, L. F. (1989). Planning for assessment: Mission statements, goals, and objectives. Trenton, NJ: New Jersey Department of Higher Education, Office of Learning Assessment.

Grandy, J. (1989). Models for developing computer-based indicators of college student learning in computer science. In Adelman, C. (Ed.), Signs and traces: Model indicators of college student learning in the disciplines, (pp. 11-64). Washington, D.C.: U.S. Government Printing Office.

Hanson, G. R. (1988). Critical issues in the assessment of value added in education. In T. W. Banta (Ed). Implementing outcomes assessment: Promise and perils. New Directions for Institutional Research, 59, San Francisco: Jossey-Bass.

Imai, M. (1986). Kaizen: The key to Japan's competitive success. New York: McGraw-Hill Publishing Company.

Learning Research and Development Center. (1990). Setting a standard: Toward an examination system for the United States. Pittsburgh: Author.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. Psychological Bulletin, 68(5), 304-305.

Miller, M. A. (1991). Assessment in hard times. Assessment Update, 3(6), 1,3,5.

Miller, G. A., & Gildea, P. M. (1987). How children learn words. Scientific American, 257(3), 94-99.

Moss, P. A. & Koziol, S. M. (1991). Investigating the validity of a locally developed critical thinking test. Educational Measurement: Issues and Practice, 10(3), 17-22.

Nickerson, R. S. (1989). New directions in educational assessment. Educational Researcher, 18(9), 3-7.

Pace, C. R. (1990). The undergraduates: A report of their activities and progress in college in the 1980s. Los Angeles: Center for the Study of Evaluation, University of California.

Pascarella, E. T. & Terenzini, P. T. (1991). How college affects students. San Francisco: Jossey-Bass.

Perkins, D. N. & Salomon, G. (1988). Teaching for transfer. Educational Leadership, 46(1), 22-32.

Perkins, D. N. & Salomon, G. (1989). Are cognitive skills context-bound? Educational Researcher, 18(1), 16-25.

Peterson, G. W. (1982). A meta-evaluation of a generic skills approach to the evaluation of academic programs. (ERIC Document Reproduction Service No. ED 219-398).

Peterson, G. W. & Hayward, P. C. (1989). Model indicators of student learning in undergraduate biology. In C. Adelman (Ed.), Signs and traces: Model indicators of college student learning in the disciplines, (pp. 93-121). Washington, D.C.: U.S. Government Printing Office.

Peterson, M. W. & Cameron, K. (1988). Organizational climate for teaching and learning. In M. W. Peterson, K. S. Cameron, A. Knapp, M. G. Spencer, & T. H. White (Eds.), Assessing the organizational and administrative context for teaching and learning: An institutional self-study manual (Technical Report No. 91-E-002). Ann Arbor, MI: National Center for Research to Improve Postsecondary Teaching and Learning.

Pike, G. R. (1989). A comparison of the College Outcome Measures Program (COMP) and the Collegiate Assessment of Academic Proficiency (CAAP) exams. Unpublished research report. Center for Assessment Research and Development, University of Tennessee, Knoxville.

Pike. G. R. (1990). Comparison of ACT COMP and College BASE. Unpublished research report. Center for Assessment Research and Development, University of Tennesse, Knoxville.

Pike, G. R. (1991a). Comparison of ACT COMP and the Academic Profile. Unpublished research report. Center for Assessment Research and Development, University of Tennessee, Knoxville.

Pike, G. R. (in press). Lies, damn lies, and statistics revisited: A comparison of three methods of representing change. Research in Higher Education.

Pike, G. R. (1991b). Using mixed-effect structural equation models to study student growth and development. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Ratcliff, J. L. (1988). Developing a cluster-analytic model for identifying coursework patterns associated with general learned abilities of college students. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Raudenbush, S. W. & Bryk, A. S. (1989). Methodological advances in analyzing the effects of schools and classrooms on student learning. Review of Research in Education, 15, 423-475.

Sternberg, R. J. (1985a). Teaching critical thinking, Part 1: Are we making critical mistakes? Phi Delta Kappan, 67(3), 194-198.

Sternberg, R. J. (1985b). Teaching critical thinking, Part 2: Possible solutions. Phi Delta Kappan, 67(4), 277-278.

Stone, H. L. & Meyer, T. C. (1989). Developing an ability-based assessment program in the continuum of medical education. Madison, WI: University of Wisconsin-Madison Medical School.

Tennessee Higher Education Commission (1977). The competent college student: An essay on the objectives and quality of higher education. Nashville, TN: Author.

Warren, J. (1984). The blind alley of value added. AAHE Bulletin, 37(1), 10-13.

Warren, J. (1989). A model for assessing undergraduate learning in mechanical engineering. In C. Adelman (Ed.), Signs and traces: Model indicators of college student learning in the disciplines, (pp. 65-91). Washington, D.C.: U.S. Government Printing Office.

Wiggins, G. (1990). The truth may make you free, but the test may keep you imprisoned. Assessment 1990: Understanding the implications (pp.17-31). Washington, D.C. American Association for Higher Education.

Wiggins, G. (1989). Teaching to the (authentic) test. Educational Leadership, 46(7), 41-47.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. Educational Measurement: Issues and Practice, 10(3), 10-16.

Paper: **Toward a Plan for Using National Assessment to Ensure Continuous Improvement of Higher Education**

Author: Trudy W. Banta, University of Tennessee-Knoxville

Reviewer: Nancy Beck, Educational Testing Service

> This is a well written, well-organized paper that outlines a national assessment effort which would involve every student and every institution in the country. The proposal is immense in its cost implications and, if by some modern miracle it were ever adopted and implemented, in its educational implications.

The paper is organized in two parts. The first section develops a very convincing case against attempting any meaningful national assessment of the 5.5 skills of critical thinking, communication, and problem solving. The politics and realities of higher education are cited as insurmountable obstacles to any hope of success. Having done that, the author then describes an elaborate system to do the impossible. It was too late; she proved so convincing in the case *against*, that I was not persuaded in the case *for* that the proposed system would or could overcome the barriers.

Banta approaches this paper from the perspective of an institutional assessment person interested in using assessment to improve curriculum, instruction, and learning. While that might be the long term aim, Goal 5 and Objective 5.5 really address results not process. It seems doubtful that policy-makers would or should share all the assumptions necessary to acceptance of this plan. Is it really necessary, for example, that *"all faculty members"* subscribe to the national goals and objectives, accept the definitions of the skills, identify all courses in which they are taught, and then teach them in ways that will enhance student learning? That would be nice, but it won't happen.

Banta acknowledges that she is writing entirely from her unique institutional experience and that gives her comments the strength of being based on the realities of the trenches - experience in a university system that was a pioneer and is a continuing force in large scale assessment. The other side of the coin, however, is that she relies on a limited University of Tennessee research base for many of her practical (as opposed to theoretical) examples. In evaluating the currently available instruments, for instance, she cites only UTK research when a broader research base is available. One example: there is research that supports the multi-dimensionality of the Academic Profile in terms of its factor structure.

When Banta describes an assessment system for the nation it reflects her own orientation and grows out of the particular assumptions she brings to the task, that is, focused on the individual student and institution. She proposes a system to: define the 5.5 abilities, have all faculty agree on the definitions, assure that the abilities as defined are taught and taught well, develop appropriate ways to measure

the abilities, have all faculty accept the measurements as appropriate, administer them to all students in ways that engage their best efforts, and then use the results to improve the methods and materials of instruction and promote student learning. It is hard to fault any of this.

It is, however, impractical, unbelievably expensive, and overly elaborate and ambitious. It is all-encompassing - all faculty and all students at all colleges and universities in the country. It is an approach that ignores the politics and realities of higher education - the very obstacles she so effectively describes in the first section of the paper.

Getting started on constructing the proposed system calls for a considerable leap of faith. As a first step, perhaps *the* critical step, the author cites the need for consensus on an operational definition of critical thinking (which she combines with problem solving). This is where we are to start even though we had been cautioned earlier that "The possibility of gaining a national consensus on stated goals/objectives for promoting critical thinking seems virtually impossible" (p. 5), with the two year struggle of the American Philosophical Association cited as an example of the intransigence of the problem (although a group of philosophers trying to agree on a definition of anything may not be the best example to use!). Having established the near impossibility of the task, this critical activity is given to 50 state panels, working independently, to address. It is hard to feel confident that they will succeed where all others have become mired in a "conceptual swamp."

At the outcomes end of the system, the key measurement aspect is the senior thesis/project to be undertaken by every senior in every institution. Using pre-determined (by a federal panel) criteria for acceptable and unacceptable performance, each project is evaluated and a detailed review is provided. Scores are reviewed and composites are made at the institutional, state, and national level. The potential cost is immense - the inevitable consequence of assuming that national goals must be measured one student at a time.

How to go about reaching national agreement on the criteria for acceptable performance across the range and diversity of the projects possible from every graduating senior is not really addressed. It has to be at least as difficult and daunting as defining critical thinking.

Although an elaborate measurement approach is outlined, details, probably purposely, are sketchy. If the task at hand was "...to identify, define, and assess a specific set of skills which are consistent with the stated objective of national goal 5..." we can, however, evaluate the measures Banta suggests against the review criteria.

1. **A valid case was proposed for the measures.** Yes, if the underlying assumption of individual and institutional feedback and focus is accepted.

2. **Acquisition or possession of the skills can be shown.** Yes, if the approach for getting to the measures is accepted. Serious questions about the viability and practicality of that approach remain, however.

3. **Permits identification of growth or value added.** Probably not. Nothing is said about equating the senior activities, and the problems of doing so, given the likely diversity of the activities, make it unlikely to occur. Without some form of equating and/or comparability of activities across students/institutions, it is hard to see how any trend data could be established.

4. **Assessments of these skills allow for:**

   - Accurate measurement of each set of skills: Hard to say at this point. If you believe that the various panels can do what the author proposes for them and some sort of agreement can be reached, then it is possible.

   - Determination of barriers to acquisition: Yes. The institutional focus of the effort maximizes the likelihood of being able to identify barriers within and across institutions.

   - Identification of effective learning environments: Yes. As above, the institutional focus should facilitate this.

5. **Methods are practical, replicable, and complete.**

   - Derived from reliable and practical research applications. No. The bibliography is long but no evidence is cited to support the senior project approach as a reliable, valid, and practical method of getting data on these skills. Indeed, the lack of cross institutional standard setting and use of common measures is critical.

- Adaptable to a national environment or program. No. Although the whole plan is presented as doable, there is no evidence in the history of higher education to suggest that such an ambitious effort could be agreed upon, funded, or carried out.

- Requires little or no further research or testing. No. Extensive research would be required, both basic and applied, before such instruments could be developed and supported.

- Cost efficient and effective practices. No!

General Comments:

While potentially attractive from an educational point of view, the proposed system would be impossible to fund or carry out. In spite of this generally negative review, this paper could be useful to policy-makers. The author very effectively shows the implications of taking a single goal/objective at face value and carrying it to its limit. What seems more likely is that policy-makers want something that will help turn the ship while recognizing that they cannot reform higher education with one objective of an overall set of broad educational goals.

November 1991

Comments on a Position Paper

by

Trudy W. Banta

Reviewed by: Norman Frederiksen

Toward a Plan for Using National Assessment to Ensure Continuous Improvement of Higher Education

Part 1. The Problem of Assessing College Students' Abilities

The part of the National Assessment project discussed in this paper in Objective 5 of Goal 5: "The proportion of college graduates who demonstrate an advanced ability to think critically, communicate effectively, and solve problems will increase substantially'; this will be refarred to an "5.5."

This is one of many of the assessment goals; some of the others are competency in ... English, mathematics, science, history, and geography" (Goal 3) "competent in more than one language"; and "ability to reason, solve problems, apply knowledge, and write and comunicate effectively" (Goal 3, Objective 2).

Goal 3-Objective 2 is almost identical to 5.5. Apparently the governors who wrote these objectives didn't compare notes before going to press.

Banta has written five assuptions regarding a plan for using National Assessments to improve higher education. But she disagrees with all of them.

Assumption 1. The Abilities Can He Defined and Agreed Upon (p. 3). Her assumptions turn out to be the opposite of her beliefs; later she states her real belief (p. 5): "The possibility of gaining a national consensus on stated goals/objectives for promoting critical thinking seems virtually impossible." But she also says that "building that consensus is absolutely essential."

I agree with her first opinion. There is too much variability among deans and professors in different kinds of collages and universities to expect anything like a consensus on goals and objectives.

Assumption 2. The defined Abilities Will Be Taught (p.6) Banta asserts that if there was a national consensus on to what was to be taught it would be taught--but only in a few small

colleges. Such a consensus could not exist in most collages and universities, where the professers decide for themselves what they should teach. Banta acknowledges that professors choose for themselves what and how they teach, and I agree.

Assumption 3: -" Abilities Can Be Measured (p. 8). Banta describes a large number of tests that are widely used to measure abilities, and she states that ideally there should be ways to compare scores for individuals over time to assess their progress. . . . " But she concludes that "current measurement theory and its application ... are inadequate to support specific suggestions for improving studental learning...based an their scores on these instruments."

I have another assumption: most, if not all, of the tests mentioned make use of the multiple-choice format. This would limit considerably the ability of the tests to assess higher-order thinking skills; multiple-choice tests tend to assess basic skills.

If one wants to assess higher-order thinking skills, it would be best to use problems that simulate real-life problems in the relevant domain (say math), and that are of the appropriate level of difficulty.

Assumption 4. Measures of the Defined Abilities Will Be Taken Seriously by the Students (p. 12)-. Banta states that "in the absence of explicit connections between their performance on a test and their academic program, students see little need to do their beet work," If the tests are conventional multiple-choice tests, I certainly agree; scores on such tests are not likely to improve student perforance. However, it in possible to develop tests that do have instructional value and might be taken seriously by students.

Assumption 5: Asessament Will Increase Student Learning (pp. 13-14), Banta expresses concern about the "measuremont of change" problem in connetion with the assessment of learning. She states that "there in a growing concern about the vast network of testing programs ... [that] has actually been an influential factor in lowering acaciamic standards.. with an overall negative impact on teaching and learning" (p. 14).

My suggestion in that the problem could be removed, or at least alleviated, by assessing successive classes rather than the same students each year in college. This is what NAEP does, with great care that each group in representative of the population being tested--national, area, or state.

Part II- -AN Approach to the Problem of Assessing College Students' Abilities.

We seem to be getting closer; Part I was "The Problem of Assessing..." and we are now up to "An Approach to the Task of Assessing...."

My concern at this point has to do with the nature of 5.5--critical thinking, communication, and problem-solving. Just what is critical thinking,. and how does it differ from ordinary thinking? What do college students think about--how to get a better room in the dorm? Who to vote for at the next election?

What to write about in your essay on Chaucer? How to prepare for the next European history exam? How to deal with a calculus problem? How to write a letter applying for a position as an instructor in the university? Who to ask for a date? All of these and much more would have to come under thinking. How can we assess a domain as large as is implied by this paragraph?

We must narrow the picture. What we are really interested in is the influnces of college attendance on learning what colleges teach--math, science, literature, or whatever courses are taught. This would narrow the assessment problem greatly. The possibility of "using testing to improve learning" (p. 15) is mentioned, and it seems to me that the idea of "embedding assessment activities in coursework" (p. 24) is a sound idea that has been tried in the lower grades.

The teacher begins by posing a problem to students, who are encouraged to form small groups to work together. Help can be provided as needed in the form of hints, reference books, models, computers, video, teacher aides and teachers, etc., as necessary. As problems are solved, more difficult complex problems can be presented. Understanding of the domain grows as success in the earlier tasks provide a background for further learning and mastery. As the term of teaching continues, records of the performance of each student can be preserved and used as a basis for assessuent.

Such procedures have been found to produce results that are far superior to the blackboard-and-eraser lectures. (See any copy of a new journal named Interactive Learning Environment, Ablex Publishing Corporation, 355 Chestnut Street, Norwood, NJ 07648).

Banta and others support "the development of a new examination system that emphasizes alternatives to traditional multiple-choice instrumental" (p. 23). This is a recommendation that I support for use in college courses. What I would prefer to see developed are tests in the form of realistic simulations of real-life problem situations in the various disciplines. The responses might be statement of what the exminee vould do or say rather than choosing options, as in a multiple-choice test.

An example in a set of "Tests of Scientific Thinking" that was intended for graduate psychology students. One of the tents is called "Formulating Hypotheses" (FH). Each FH problem requires the examines to (1) read a brief description of an experiment; (2) study a graph or table showing the results of the experiment; (3) read a statement of the major finding; and (4) write hypotheses (possible explanations) that might account for the finding. The problem has no single right answer, but there are many, possible answers that vary widely in quality. The scoring system involves (1) making a classification of the ideas written by the students who took the test, thus forming a set of mutually exclusive categories, and (2) having the categories valued by expert judges. Scoring then involves assigning each response to one of the categories and letting the computer do the rest. (Sos Frederiksen, N., & Ward, W. 0. (1978), Measures for the study of creativity in scientific problem solving. Applied Psychological Measurement, &, 1-24; and Ward, W. C. Frederiksen, N., & Carlson, S. B, (1980). Construct validity of free-response

and machine scorable forms of a test. Journal of Education Measurement, 17, 11-29).

The scores on such tests clearly involve thinking (whether critical or not) and problem-solving. Communication in illustratad by what we wrote, I presume. Thus the demands of 5.5 have been satisfied and higher-order skills can be measured.

Review of:

Trudy Banta: "Toward a Plan For Using National Assessment to Ensure Continuous Improvement of Higher Education"

By: Barbara Wright and Ted Marchese, AAHE Assessment Forum

After setting forth five assumptions that she views as implicit in National Education Objective 5.5, and arguing that little evidence exists to support any of these assumptions, Banta goes on to propose a postsecondary assessment program that will link assessment with educational improvement. Banta's opening points strike us as generally valid though a bit overstated; in the second half of the paper, she proposes a national assessment-and-improvement program that addresses many of the problems raised in part I.

Turning first to the opening set of assumptions, we wonder whether consensus on a single definition of critical thinking et al. is really "essential" (p.5) or even desirable, much less possible. This may be the conventional wisdom when we're looking at a traditional high-stakes testing situation. But moving away from that context, it can be argued that such consensus would lead to a disastrous reductionism, a dangerous impoverishment of what we mean by "critical thinking." Doesn't such an assumption impose a kind of scientific rationalism on the chaotic richness of human life, intellectual styles, and contexts for thought? Don't we thus confuse "uniformity" with "quality"?

Of course, diversity doesn't guarantee quality, any more than uniformity does. But in an analogy to the value we place on biological diversity, in the interests of robustness, adaptability, and fairness, it makes sense to encourage or at least accommodate the widest possible range of variation in intellectual processes. The participants in this gathering can doubtless think

of many ways to handle the issue of definition to allow maximum flexibility.

Similarly, the measurement of critical thinking skills (assumption #3) can indeed be problematic, particularly if we lack consensus on definition, if we assume that measurements must be taken using the typical standardized tests and formats, and if we insist on precise, quantified results that will be reported for high-stakes purposes. But what if our purpose instead is to collect evidence that demonstrates students' ability to use critical thinking skills? This sort of approach may be unconventional in education, but it's not unheard of; and it enjoys respect elsewhere, for example, in our legal system, where human judgment must be brought to bear on complex questions of guilt or innocence, motivation, character, circumstances, and punishment or acquittal.

Another advantage of collecting evidence, as opposed to scores or in addition to scores, is that actual examples of tasks could be published, along with a range of student responses. Why do that? Apart from their value as demonstrations of accountability, such concrete examples would have educative value -- for teachers, students, parents, employers, society at large. Education in the US has been hit hard by anecdotal reports about what students don't know and can't do; positive examples of what students can do, concrete demonstrations of what critical thinking is and how it works, could go a long way toward both balancing the picture of American education and promoting wider acquisition of the skills. Each example, to mangle a metaphor, could be worth a thousand scores, not just in the classroom but beyond.

The notion that critical thinking abilities will be taught and that students will learn them (Banta's assumption #2) has been demonstrated at least in California, where over a decade ago Executive Order 338 mandated state-wide instruction in critical thinking. Tests have shown that students who took critical thinking courses did improve their abilities. (See, for

example, Nummedal, Halpern, Marsh, and Carter-Wells, "A Multidimensional Approach to the Assessment of Critical Thinking," presented at AAHE Conference on Assessment in Higher Education, San Francisco, June 10, 1991.) This was accomplished even as the different sectors -- K-12, the community colleges, the CSU system, the UC system -- were allowed to develop their own definitions. The California example suggests that the most important thing is not what is taught or how, but that critical thinking is consciously taught at all.

As for whether students will put out their best efforts (assumption #4) or whether results will be used to improve learning (assumption #5), the answer is not that this never happens or cannot, but that it depends a great deal on people and institutions, their values, resources, reward systems, and the political context in which they function. There are successful examples; the issue is whether educators and public officials choose to act on them.

The second part of the paper sets forth the dimensions of a full national program not only to measure but to improve collegiate attention to the three abilities. Significantly, Banta suggests alternatives to traditional multiple-choice instruments: theses, projects, and other capstone experiences, along with portfolios to provide a view of student development over time. Such a plan clearly responds to the difficulties raised by the five assumptions. It should not be simply dismissed as "impractical" or "too expensive"; indeed, it might be the best investment a nation could make.

The paper reminds us that it's extraordinarily difficult for any single assessment system to serve the twin masters of public reporting and data for improvement, but Banta accepts the challenge, arguing that the necessary, enormously costly investment in a national assessment system will only make

sense if accountability is part of a larger effort to improve education.

Nevertheless, given Banta's five earlier points, the lack of consensus about definitions or methods (not necessarily a bad thing, as we have argued above, but essential in her view), the need for new forms of accountability data, and the high stakes as well as the high cost involved, it seems reasonable to ask whether it wouldn't make more sense to stop short of the full proposal.

It strikes us as more do-able to adopt a slimmed-down version of the proposal. One could begin by lining up a sample of institutions (in the dozens, no more), perhaps grouped according to the definitions of critical thinking they found most congenial with their mission or institutional culture. The institutions in each group could agree to bring forward portfolios of representative work from a sample of students (seniors). The next task would be for panels of external experts to review that work and come to some judgments: Are the definitions workable? Do the portfolios bring forward the necessary evidence? What feedback to institution and to student do they provide? How valuable is it? What can be gathered from such a process by way of data, information, or examples that would have value to decision-makers and the general public?

Two or three iterations of this process might be needed to get it right -- lean yet useful, credible, flexible. At that point, it could go "public" more widely. It need not be forced upon every student at every institution every year in order to begin to be influential; we can imagine consortia or state systems of higher education adopting it because it meets real needs -- internal feedback and consciousness-raising, along with a credible way to speak to the public about issues of institutional and student performance. Though we don't necessarily endorse this, it's easy to foresee

such a system, once proven, becoming an object of state mandate and an accredition requirement. And once such a system for public reporting catches hold, it surely will raise demand for the good practices and faculty development described in this paper.

In the end, 10 years from now, these two approaches -- the one we envision here and that advanced in this paper -- may bring us to the same point: wide acceptance and pursuit of the three abilities. The difference is between a top-down, national, all-at-once approach and a more evolutionary, developmental, flexible one. First, let's see whether we can operationalize the thing, then set loose engines for its adoption.